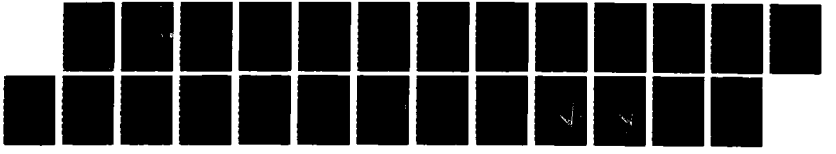
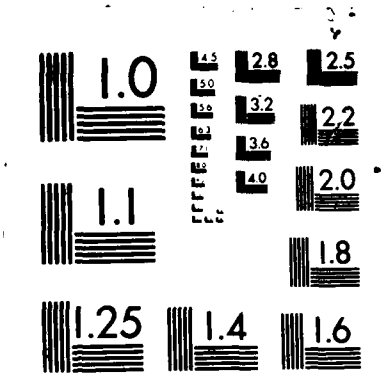
AD-A183 326 OPTIMIZATION IN ANALYTICAL CHEMISTRY USING ROBUST ETTIMATION(U) UTAH UNIV SALT LAKE CITY DEPT OF CHEMISTRY G R PHILLIPS ET AL 30 JUL 87 TR-2 1/1

UNCLASSIFIED N00014-86-K-0639 F/G 12/3 NL





OFFICE OF NAVAL RESEARCH

Contract N00014-86-K-0639

R & T Code 4133013---02

Technical Report No. 2

Optimization in Analytical Chemistry Using Robust Estimation

by

Gregory R. Phillips and Edward M. Eyring

Prepared for Publication

in

Analytical Chemistry

University of Utah
Department of Chemistry
Salt Lake City, UT

July 30, 1987

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE

AD-A183 326

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release and sale. Distribution unlimited
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) ONR Technical Report No. 2	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION University of Utah 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION Office of Naval Research Resident Representative
6c. ADDRESS (City, State, and ZIP Code) Department of Chemistry University of Utah Salt Lake City, UT 84112	7b. ADDRESS (City, State, and ZIP Code) University of New Mexico Bandelier Hall West Albuquerque, NM 87131
8a. NAME OF FUNDING / SPONSORING ORGANIZATION Office of Naval Research 8b. OFFICE SYMBOL (If applicable) ONR	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-86-K-0639
8c. ADDRESS (City, State, and ZIP Code) 800 N. Qunicy Street Arlington, VA 22217	10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO

11 TITLE (Include Security Classification)
Optimization in Analytical Chemistry Using Robust Estimation

12 PERSONAL AUTHOR(S)
Gregory R. Phillips and Edward M. Eyring

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical	FROM _____ TO _____	1987, July 30	22

16. SUPPLEMENTARY NOTATION

17 COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Robust estimation, least squares estimator, Huber estimator, Gaussian distribution, non-Gaussian error distribution

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

Analytical chemists have long been concerned with obtaining optimal experimental conditio Robust estimation provides an additional method of increasing the efficiency of an analytica technique. This is illustrated for the determination of the "true" value, μ, of a quantity which is measured with error. The least squares estimator of μ is compared with the median and Huber estimates over a variety of error distributions in the vicinity of the Gaussian distribution. Simulation allows examination of the efficiency of an estimation procedure as a function of the error distribution. Results are presented which show the least squares estimator of μ to be much more sensitive to a non-Gaussian error distribution than generally realized in the chemical community. Additionally, the arguments commonly used to support least squares estimation are critically examined.

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21 ABSTRACT SECURITY CLASSIFICATION Unclassified
22a NAME OF RESPONSIBLE INDIVIDUAL	22b TELEPHONE (Include Area Code) / 22c OFFICE SYMBOL

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted — SECURITY CLASSIFICATION OF THIS PAGE

OPTIMIZATION IN ANALYTICAL CHEMISTRY USING ROBUST ESTIMATION

Gregory R. Phillips and Edward M. Eyring*
Department of Chemistry
University of Utah
Salt Lake City, Utah 84112
(801) 581-8658

ABSTRACT

Analytical chemists have long been concerned with obtaining optimal experimental conditions. Robust estimation provides an additional method of increasing the efficiency of an analytical technique. This is illustrated for the determination of the "true" value, μ, of a quantity which is measured with error. The least squares estimator of μ is compared with the median and Huber estimates over a variety of error distributions in the vicinity of the Gaussian distribution. Simulation allows examination of the efficiency of an estimation procedure as a function of the error distribution. Results are presented which show the least squares estimator of μ to be much more sensitive to a non-Gaussian error distribution than generally realized in the chemical community. Additionally, the arguments commonly used to support least squares estimation are critically examined.

INTRODUCTION

Experimental optimization has been an important subject in analytical chemistry for many years now. This term often, though not always, suggests a technique for increasing the precision of analytical measurements (e.g. increased sensitivity, improved reliability, or decreased cost). Examples of optimization in chemistry range from the development of self-optimizing instruments(1) to the use of expert systems in methods development(2).

The efficiency of an analytical technique depends on more than just the precision of the measurement process. Eckschlager and Stepanek(3) have characterized an analytical system as two relatively independent subsystems. In the first of these two subsystems, an analytical apparatus extracts information from a sample and encodes it in an analytical signal (e.g. voltage); in the second, this signal is decoded to yield information. The information gained from a chemical analysis depends on the efficiency of the overall system, and can be limited by either of the two subsystems. Most of the optimization done in analytical chemistry has been concerned with the first subsystem.

The problem of decoding analytical signals lies within the realm of chemometrics, which has been defined as the discipline of using mathematical and statistical techniques to extract information from measurements(4). Chemists often associate chemometrics with sophisticated multidimensional techniques, expert systems, or artificial intelligence. In spite of very elegant work in these areas, the vast majority of chemometric techniques actually used in chemical laboratories are simple univariate statistics, such as least squares estimates of the mean, standard deviation, or regression coefficients. These statistics are usually justified in analytical texts by the assumption of Gaussian, or normal, errors.

The importance of the normal error distribution to least squares techniques, along with the consequences of departures from this assumption, has received much attention from statisticians; however, most chemists seem to be largely unaware of its importance. Ames and Szonyi(5) and Filliben(6) have warned of the possibility of drawing incorrect conclusions when the normality assumption is violated, and have proposed the testing of error distributions. Tests for normality require many more observations than are generally available in chemical experiments. Even when an adequate number of data points is available, it is most unusual for a chemist to apply any normality test. Studies in enzyme kinetics have both supported(7,8) and contradicted(9-11) the assumption of normal error distributions in chemical data. In a particularly impressive study, Clancy(12) has examined 250 error distributions based on 50,000 chemical analyses and found less than 15% of the distributions can be considered normal for the purpose of applying common statistical techniques.

Many statistics books for the research worker deal exclusively with least squares methods, and only invoke the assumption of independent, normally distributed errors for the validity of confidence intervals and statistical tests calculated using least squares results. Thus it is not surprising that many chemists believe least squares estimates are the optimum statistics whatever the error distribution. The efficiency of these estimates rapidly decreases under mild departures from normality, as has been demonstrated by several recent studies and is discussed in further detail below. In terminology familiar to the analytical chemist, nonnormal errors can lead to poor precision in least squares parameter estimates and inaccuracy in statistical tests and confidence intervals.

Much work is currently underway in statistics in the development of robust estimation, as illustrated by references 13-15. A statistic is called robust if it is insensitive to mild departures from the underlying assumptions and is only slightly inefficient relative to least squares when these assumptions are true. This inefficiency under ideal circumstances is often referred to as the premium paid for protection under nonideal conditions. Additionally, robust methods are also resistant to the presence of any outliers in the data. Unlike statisticians, chemists have paid only passing attention to these developments. Isenburg(16) has proposed the method of moments as an alternative to least squares iterative reconvolution in the analysis of pulse fluorometric data. Phillips and Eyring(17) and Massart et al.(18) have compared the performance of least squares regression and robust regression, concluding that robust regression often outperforms least squares regression in the analysis of chemical data. The main emphasis behind these articles has been the insensitivity of robust estimation to a small number of errors in the data.

The present paper is concerned with robust estimation as a method of increasing the efficiency of an analytical technique. This can best be illustrated by the estimation of the "true" value, μ, of a quantity which is measured with error. For example, this may be the concentration of Pb in drinking water. The least squares estimator of μ is compared with robust estimates over a variety of realistic error distributions in chemistry. Simulation allows examination of the efficiency of an estimation procedure as a function of the error distribution. Additionaly, the arguments commonly used to support least squares estimation are critically examined.

EXPERIMENTAL

Robust estimation The least squares estimate of μ is the arithmetic mean. This is often denoted by $\bar{x}$ and referred to simply as the mean. A robust estimate of μ can be obtained from the weighted mean of the observations using the Huber weight function. This is not the only method of robust estimation, nor necessarily the best, but will serve to illustrate the potential advantages of robust estimation. This approach is also conceptually simple and easy to implement.

$$\overset{r}{x} = \frac{\Sigma_1^n w_i x_i}{n} \tag{1}$$

Huber's weight function is defined by

$$w_i = \begin{cases} 1 & |r| < kS \\ (kS)/|r| & |r| > kS \end{cases} \tag{2}$$

where r is the residual (i.e., difference between observed and predicted responses), k (the tuning constant) determines how harshly large residuals are treated, and S is an estimate of the standard deviation. The evaluation of weights requires an estimate of μ. The initial estimate used in the present work is the median.

The most common measure of standard deviation is the root mean square of the residuals. This is the optimal estimator for a normal error distribution, but rapidly loses its advantages over other estimators under even slight deviations from normality(19). Additionally, a single large residual can drastically change the value of the estimator. The measure of

standard deviation used in the present work is the normalized median of the absolute deviations:

$$S = 1.48 \cdot \text{median}\{|r_i|\} \qquad (3)$$

Figure 1 shows a graph of Huber's weight with k=1.5 as a function of the residual normalized by the standard deviation. Observations within 1.5 standard deviations of the predicted value receive full weight. (For a normal error distribution, 87% of the errors fall in this region(20).) Observations outside this range receive smaller weights as they become less consistent with the remaining observations. The choice of a value for k is a compromise between two opposing tendencies: smaller values of k are more efficient for non-Gaussian errors, but less efficient when the errors are actually from a Gaussian distribution(14).

Simulated Data Four hundred different error distributions were simulated on a VAX 8300 computer. Each distribution is a combination of two Gaussian distributions:

$$\varepsilon = (1-\alpha)\ N(0,1) + \alpha\ C\ N(0,1) \qquad (4)$$

where α is the probability of contamination, C is the degree of contamination, and N(0,1) denotes the standard normal error distribution. These error distributions, referred to as contaminated normals, are a mixture of observations from a normal error distribution with $\sigma = 1$ with probability $1-\alpha$ and from an error distribution with $\sigma = C$ with a probability of α. Values of C less than, equal to, and greater than one correspond to error distributions narrower than, identical to, and wider

than the standard normal (i.e. Gaussian) distribution. This work used values in the range $0 < \alpha < 0.20$ and $0 < C < 6$. Figure 2 presents the ideal error distribution and the most extreme distribution used. The use of the standard normal as a reference is completely general and does not affect the conclusions reached.

Gaussian errors were generated by combining the methods of Wichmann and Hill(21) and Beasly and Springer(22). Three simple multiplicative congruential generators produce numbers uniformly distributed between 0 and 1. These random numbers are transformed into normal random deviates by the method of Beasly and Springer. Both algorithms are written in FORTRAN, and are machine-independent. A histogram of 1000 simulated errors is shown in Figure 3, along with the theoretical distribution. Agreement between the two is excellent.

RESULTS

This paper considers three statistics, each of which is a valid estimator of μ. However, each statistic is not equally effective in extracting the information encoded in analytical signals. Each estimator is a function of several random variables, and is therefore a random variable itself. By repeatedly simulating sets of "experimental measurements", it is possible to generate the distribution of the estimates themselves.

For each error distribution, 5000 simulated data sets (each containing 10 observations) were analyzed by the arithmetic mean, median, and H15 estimators. (H15 is shorthand notation for the weighted mean using Huber weights with $k = 1.5$.) The variance of each procedure was evaluated for

each error distribution (i.e. each combination of α and C). For example, the variance of the arithmetic mean is given by

$$Var(mean;\alpha,C) = \Sigma_1^{5000}(\bar{x}_i - \mu)^2 / 5000 \qquad (5)$$

The efficiency of the Huber and median estimators are defined relative to the arithmetic mean by

$$Eff(H15;\alpha,C) = Var(mean;\alpha,C)/Var(H15;\alpha,C) \qquad (6)$$

$$Eff(median;\alpha,C) = Var(mean;\alpha,C)/Var(median;\alpha,C) \qquad (7)$$

The relative efficiencies of the H15 and median estimators are shown in Figures 4 and 5, respectively. The increase in precision is particluarly dramatic when the narrow range of distributions studied around exact Gaussian errors (see Figure 2) is considered. Each error distribution studied was "close" to Gaussian and symmetric. Intoduction of asymmetry would have further deteriorated the precision of the mean(14).

The relative efficiency measures the precision of an estimator, such as the Huber or median, relative to the mean for the same number of observations. For an ideal Gaussian error distribution, the relative efficiencies of the H15 and median estimators would be ~.95 and .67, respectively. Under the most exterme conditions studied in this work, the relative efficiency of the H15 and median were 3.25 and 2.73. Thus, the variance of the estimated value of μ using the arithmetic mean is 3.25 times that of the Huber estimator, on the average.

Figure 6 shows a contour plot of the relative efficiency of the H15 estimator as a function of the probability of contamination and degree of contamination. Dashed contours denote regions where the arithmetic mean is more precise, while solid lines denote regions where the H15 estimator is more precise. In view of the greatly enhanced precision of robust estimation under slight deviations from normality, the small premium under ideal conditions appears quite worth the improved efficiency of robust estimation under nonideal conditions.

DISCUSSION

The prevalent attitude among chemists seems to be that rejection of erratic data points provides sufficient protection against nonnormal error distributions and justifies the automatic use of least squares procedures. The reasons given in support of least squares estimators deserve examination. Least squares statistics are easy to compute; in fact, this was one reason for the historical acceptance of least squares. However, with the proliferation of laboratory microcomputers, or even pocket calculators, ease of computation is no longer of primary importance.

A second reason for the widespread belief in least squares is a result of a misintrepretation of the Gauss-Markov theorem(23). This theorem states that the best linear unbiased estimate of μ is the sample mean, whatever the error distribution. This is frequently intrepreted by nonstatisticians to mean that the sample mean is the best of all estimators. The important words in the Gauss-Markov theorem are linear and unbiased. A linear estimator is one which is a linear combination of the observed values. However, their is no inherent reason to require

linearity. As has been shown, insistence on linearity can result in a loss of precision.

Since least squares is the optimum estimation procedure for normally distributed errors, a third argument is that it should be almost optimum when the errors are approximately normal. The Central Limit Theorem states that the sum of a "large" number of independent random variables (i.e., errors) is approximately normal regardless of the distribution of the individual random variables(20). Experimental errors are the sum of many small independent errors. However, these small errors often have widely different variances and the "approximately" normal distribution of their sum is closer to a long-tailed distribution. Studies over the past 15 years have shown the arithmetic mean to be significantly less efficient in these situations. The error distributions used in this work have only slightly longer tails than the normal distribution, yet clearly demonstrate the loss of precision in the arithmetic mean.

Finally, it is interesting to compare the present relationship between the arithmetic mean and the normal error distribution with the historical relationship. Gauss(23) introduced the normal, or Gaussian, error distribution in 1821. He argued that it was impossible to determine the most probable value of an unknown quantity unless its error distribution was known. Without such knowledge, the only recourse was to assume a distribution in a "hypothetical" fashion. Gauss preferred to take the opposite approach and to look for that distribution which would make the arithmetic mean the best estimator. Thus, the arithmetic mean was used to justify the normal error distribution.

The method of least squares has proven very useful for many years. This procedure is often motivated as being the maximum likelihood estimator

for a Gaussian error distribution. Methods for robust estimation do not represent an abandonment of traditional data reduction procedures. Estimation using robust weights is attractive since it represents the maximum likelihood estimator over a range of distributions in the "vicinity" of Gaussian. Thus, the attractive features of the Huber estimator do not depend on the existence of an idealized error distribution.

CONCLUSION

Techniques based on the principle of least squares are the optimal estimation procedures for the analysis of data possessing a normal error distribution, but perform very poorly in situations involving a nonnormal error distribution (see, e.g., reference 14). Almost every aspect of the measurement process has been examined during optimization procedures. However the validity of the assumption of normal errors has received little attention from chemists. The present work has demonstrated that even small deviations from normality can seriously degrade the efficiency of least squares estimators. Only symmetric error distributions have been examined here (more serious problems arise when the error distribution becomes asymmetric.) The deviations are so small as to frequently occur in practice. The effect of this can be to decrease the precision of an analytical method or instrument which has been carefully optimized.

Robust estimation is a complementary technique which is relatively efficient over a broad range of error distributions. This approach takes advantage of the "a priori" knowledge that errors in chemistry lie within a range of distributions, while avoiding the inefficiency which results from rigid assumptions about the error distribution. These procedures more

closely reflect real situations, recognizing that even in careful work the distribution of errors is not always ideal. Robust procedures do not change the focus of data analysis, rather they are an efficient alternate method of accomplishing traditional goals. The exact robust procedure used is not as important as the use of some robust method. This can be a newer robust approach, such as the Huber weight function, or a more traditional method of examining the validity of least sqaures.

Robust methods should not be regarded as a completely automatic procedure or a substitute for a reasonable amount of statistical knowledge, however. Measurements which have been assigned small robust weights have been marked for special attention, including examination of the appropriateness of the error model as well as the possibility of erroneous data points.

It is not the contention of this paper that improved statistical techniques, such as robust estimation, are a substitute for good analytical data. No statistical technique can extract high quality results from low quality data. If the measurement process is not in control, an analyst will benefit most by restoring the experimental conditions to their optimum values. Conversely, when a measurement process is in control, analytical precision can be limited by application of inefficient statistical procedures. Robust estimation is one method of detecting incorrect statistical models and/or error distributions. It has the advantages of being easily implemented and understood.

LITERATURE CITED

1. Deming, S.N.; Parker, L.R. CRC Critical Reviews in Analytical Chemistry 1978, 7, 187.

2. Bach, R.B.; Karnicky, J.; Abbott, S. In "Artificial Intelligience Applications in Chemistry"; T.H. Pierce and B.A. Hohne, Eds.; American Chemical Society Symposium, Ser. 306, American Chemical Society: Washington, D.C., 1985.

3. Eckschlager, K.; Stepank, V. Anal. Chem. 1982, 54, 1115A.

4. Sharaf, M.A.; Illman, D.L.; Kowalski, B.R. "Chemometrics"; Wiley: New York, 1986.

5. Ames, A.E.; Szonyi, G. In "Chemometrics: Theory and Applications"; Kowalski, B.R., Ed.; American Chemical Society Symposium, Ser. 52, American Chemical Society: Washington, D.C., 1977; pp 219-242.

6. Filliben, J.J. In "Validation of the Measurement Process"; DeVoe, J.R., Ed.; American Chemical Society Symposium, Ser. 63, American Chemical Society: Washington, D.C. 1977; pp 30-113.

7. Siano, D.B.; Zyskind, J.W.; Fromm, H.J. Arch. Biochem. Biophys. 1975, 170, 587.

8. Askelof, P.; Korsfeldt, M.; Mannervik, B. Eur. J. Biochem. 1976, 69, 61.

9. Nimmo, I.A.; Mabood, S.F. Anal. Biochem. 1979, 94, 265.

10. Storer, A.C.; Darlison, M.G.; Cornish-Bowden, A. Biochem. J. 1975, 151, 361.

11. Mabood, S.F.; Newman, P.F.J.; Nimmo, I.A. Biochem. Soc. Trans. 1977, 5, 1540.

12. Clancy, V.J. Nature 1947, 157, 339.

13. Hoaglin, D.C.; Mosteller, F.; Tukey, J.W., Eds. "Understanding Robust and Exploratory Data Analysis"; Wiley: New York, 1983.

14. Andrews, D.F.; Bickel, P.J.; Hampel, F.R.; Huber, P.J.; Rogers, W.H.; Tukey, J.W. "Robust Estimates of Location: Survey and Advances"; Princeton University Press: Princeton, NJ, 1972.

15. Huber. P.J. "Robust Statistics"; Wiley: New York, 1981.

16. Isenburg, I. Biophys. J. 1983, 43, 141-148.

17. Phillips, G.R.; Eyring, E.M. Anal. Chem. 1983, 55, 1134.

18. Massart, D.L.; Kaufman, L.; Rousseeuw, P.J.; Leroy, A. Anal. Chim. Acta 1986, 187, 171-179.

19. Fenstad, G.U.; Kjaernes, M.; Walloe, L. J. Statist. Comput. Simul. 1980, 10, 113-132.

20. Dwass, M. "Probability and Statistics"; Benjamin: New York, 1970.

21. Wichmann, B.A.; Hill, I.D. Applied Stat. 1982, 31, 188-190.

22. Beasley, J.D.; Springer, S.G. Applied Stat. 1977, 26, 118-121.

23. Huber, P.J. Annals of Math. Statist. 1972, 43, 1041-1067.

CREDIT

This research was supported in part by the Office of Naval Research. The computer was funded in part by National Science Foundation Grant CHE 8416337.

Figure Captions

Figure 1. Plot of the Huber weight function with a tuning constant equal to 1.5. The dashed line is the probability density for the Gaussian error function.

Figure 2. A plot of Gaussian error distribution (_____) and a contaminated distribution with $\alpha = 0.20$ and $C = 6$ (__ __).

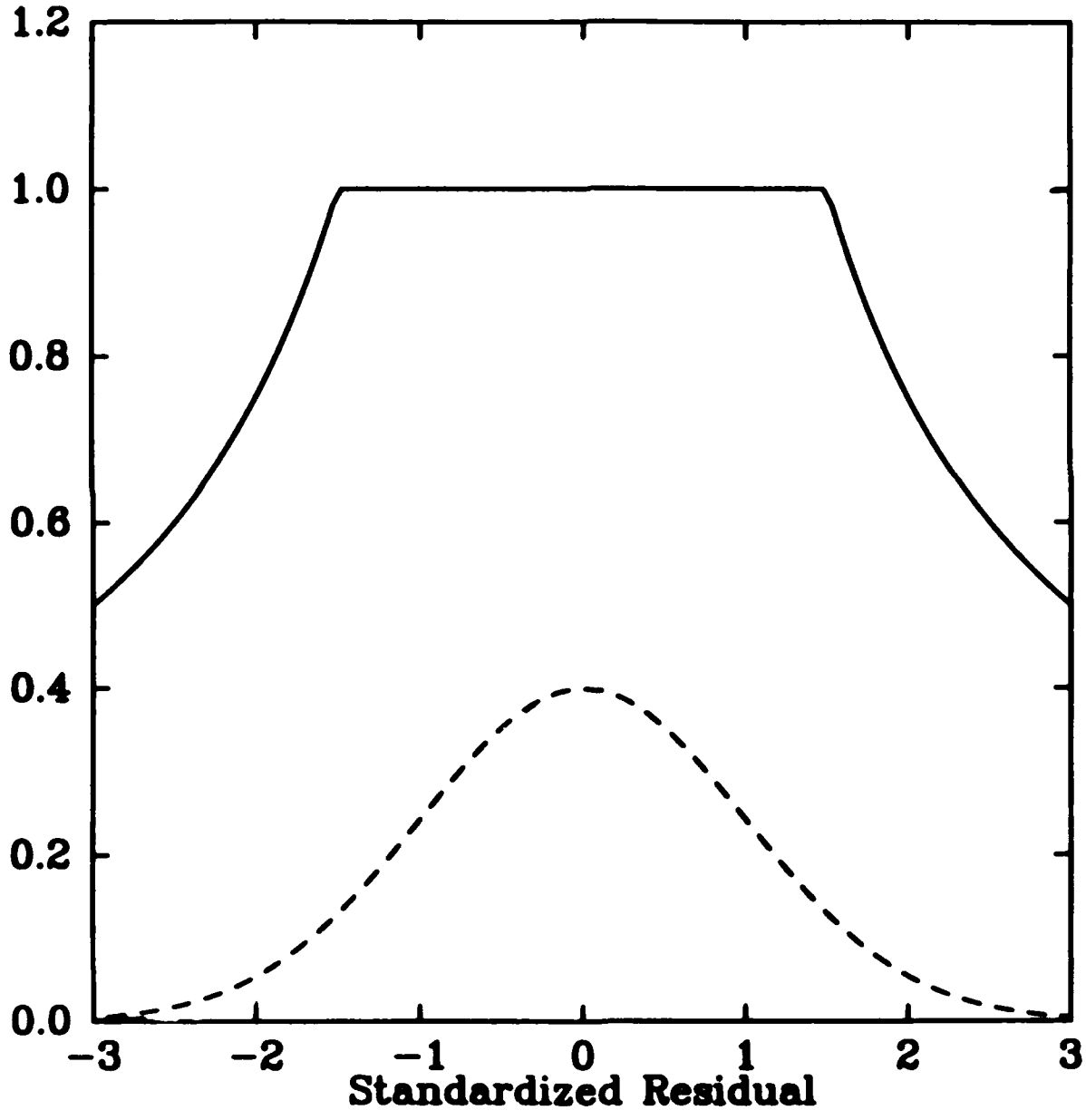
Figure 3. Histogram of 1000 simulated errors. Superimposed is the theoretical distribution for normal, or Gaussian, errors.

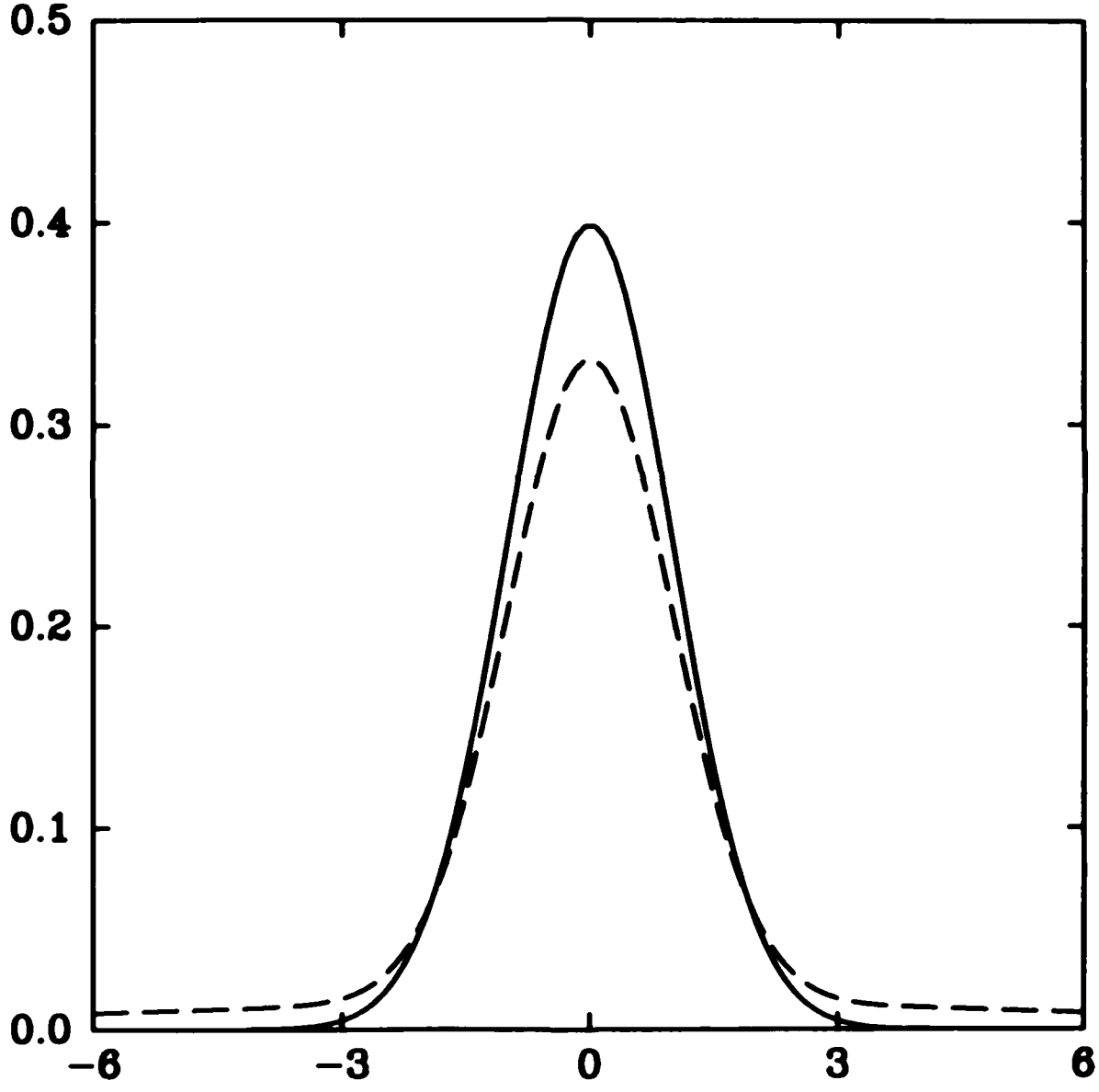
Figure 4. The relative efficiency of robust estimation using the Huber weight function with k=1.5 as function of the probability of contamination, ε, and the degree of contmaination, C.

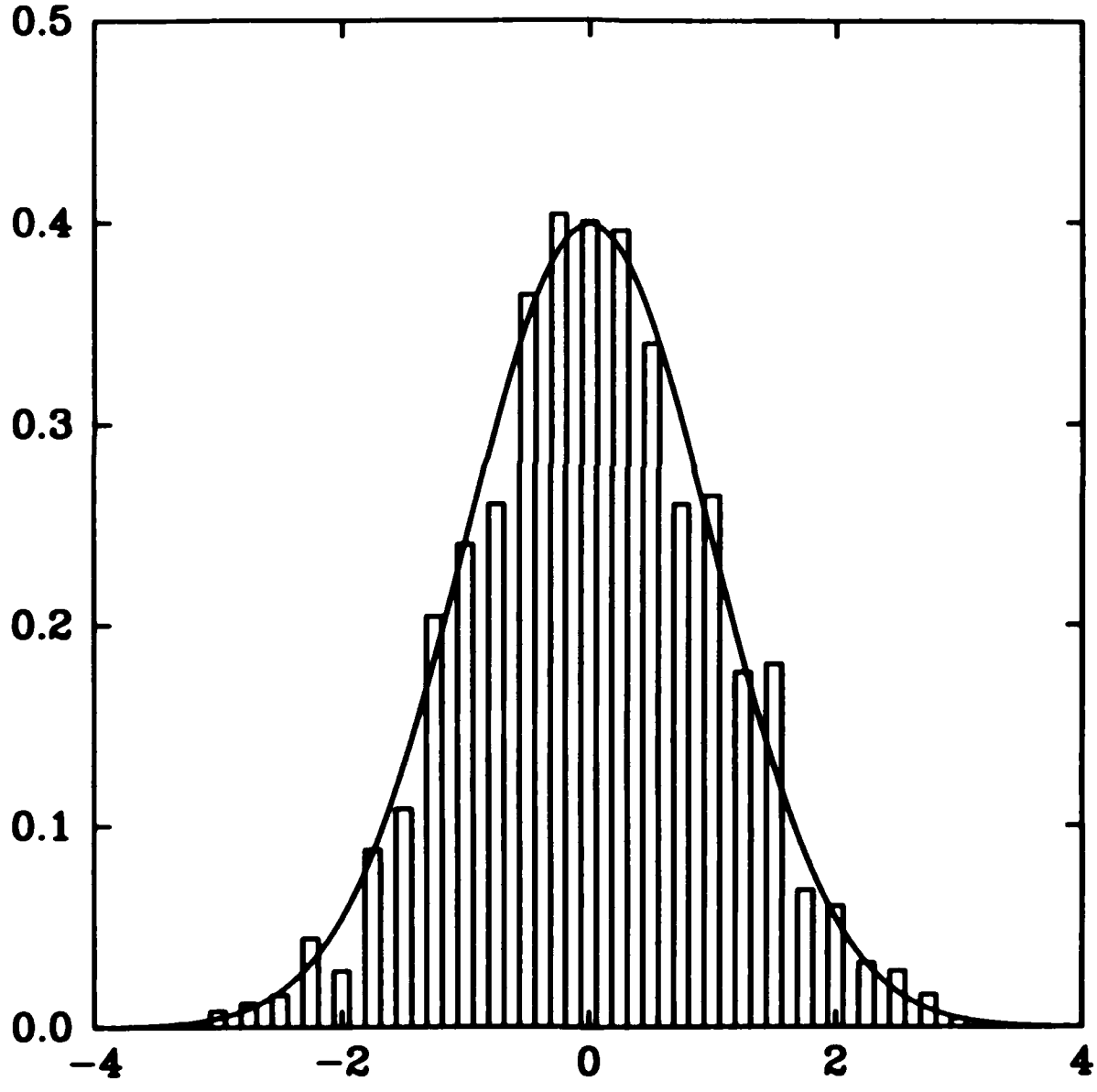
Figure 5. The relative efficiency of the median estimator as function of the probability of contamination, ε, and the degree of contmaination, C.

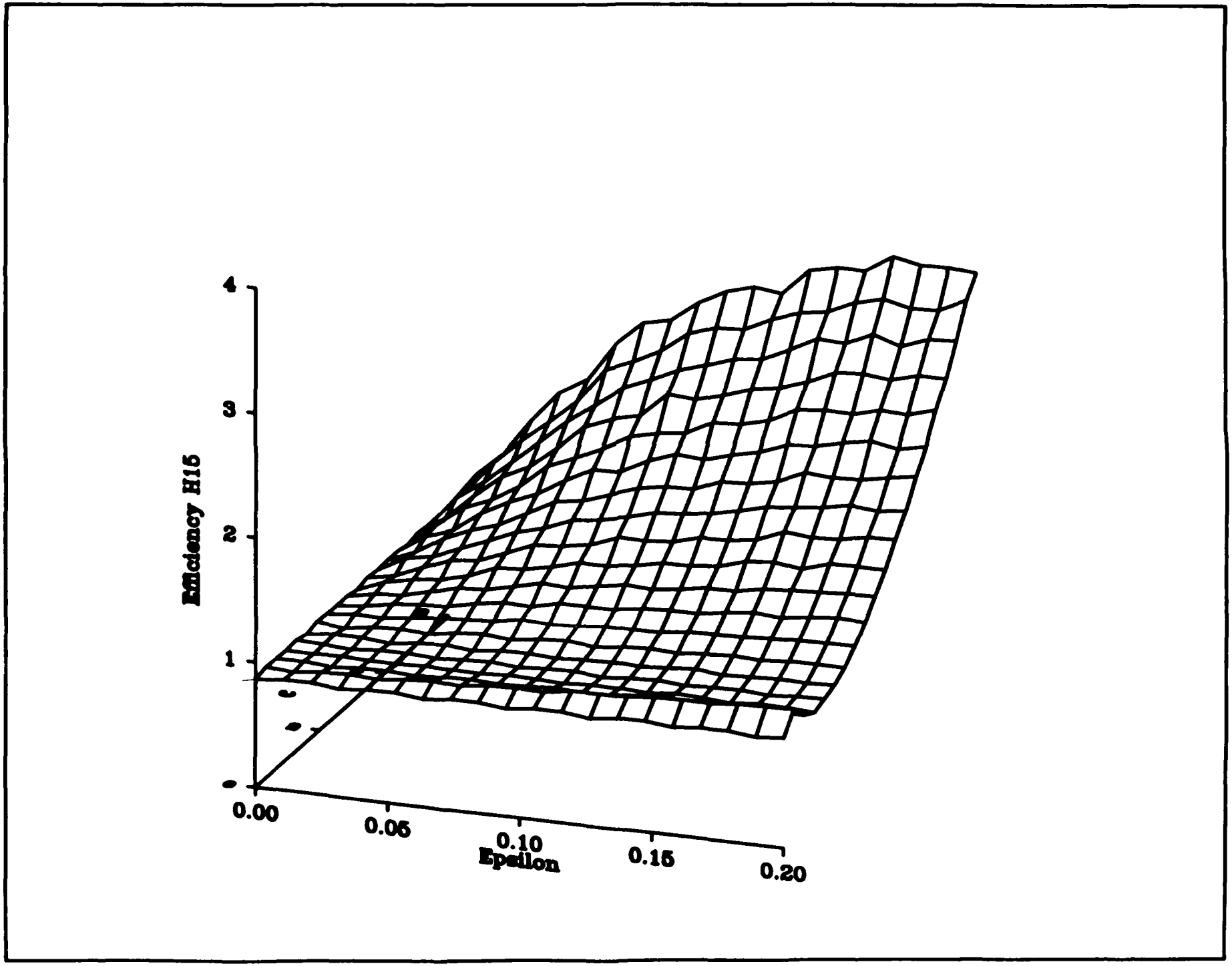
Figure 6. A contour plot of the relative efficiency of the H15 estimator on contaminated normals. Dashed lines correspond to efficiencies less than one; solid lines correspond to efficiencies greater than one.

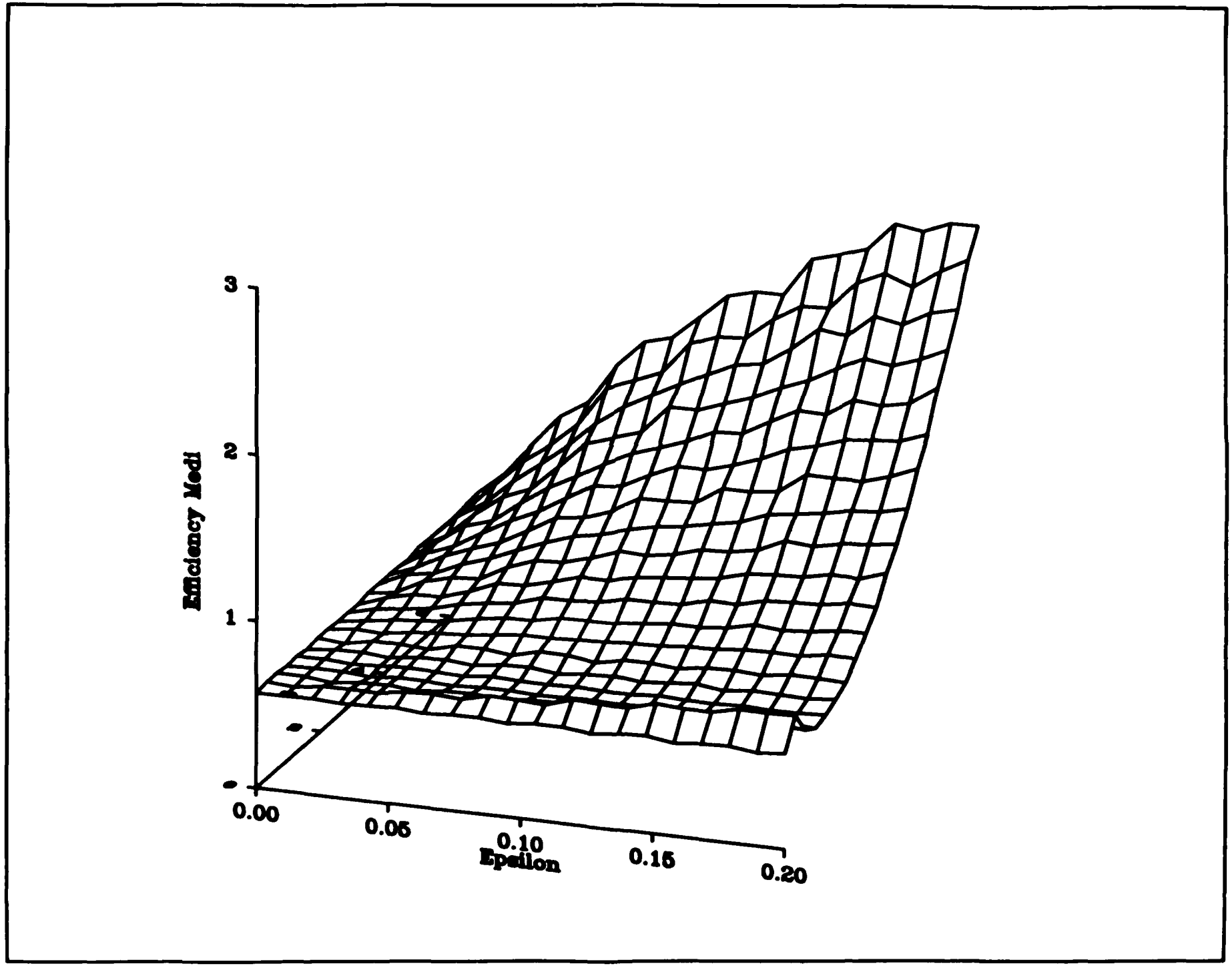
1.5 Huber

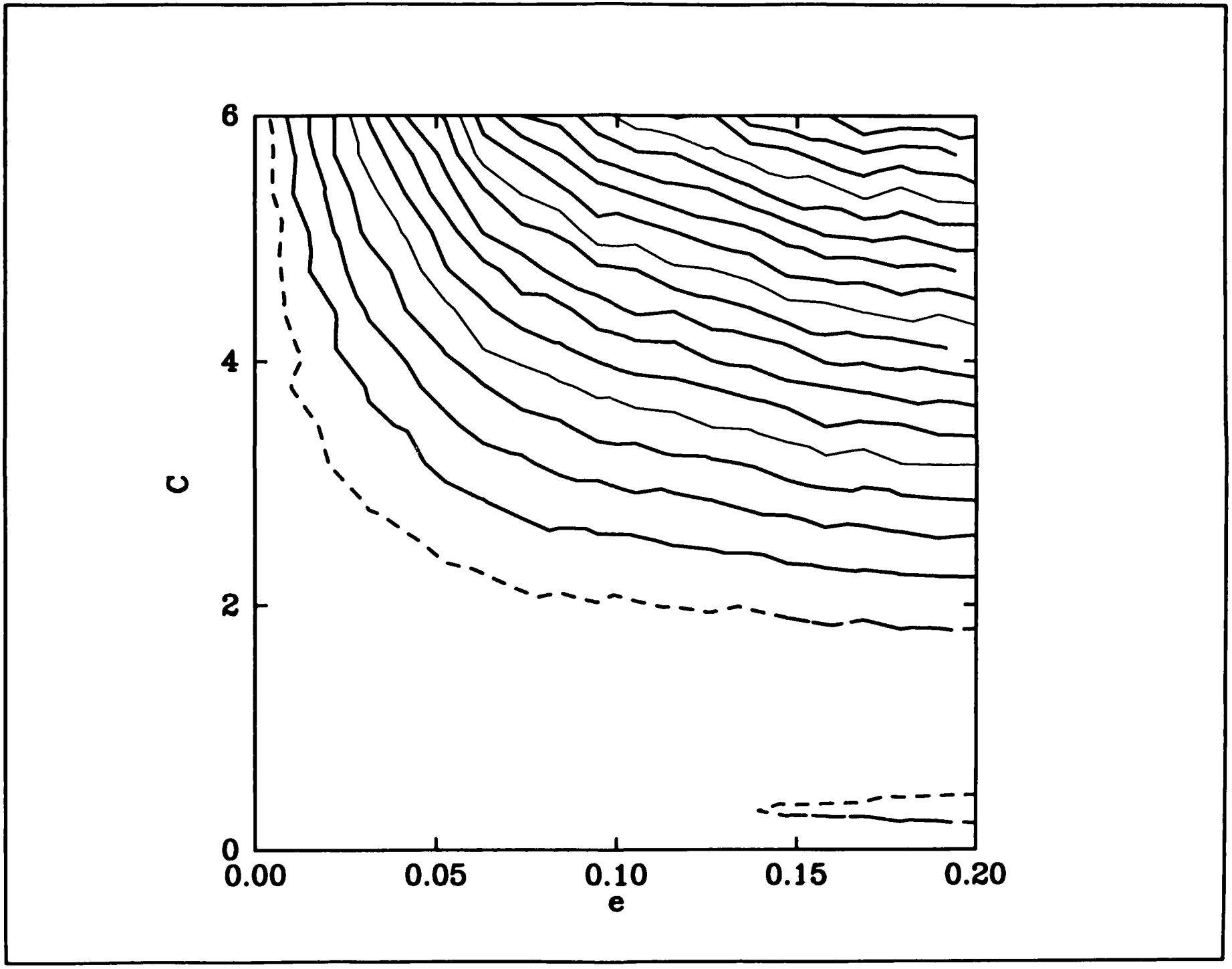



0.5
0.4
0.3
0.2
0.1
0.0
−6
−3
0
3
6


0.5
0.4
0.3
0.2
0.1
0.0
−4
−2
0
2
4


Efficiency H15
4
3
2
1
0
0.00
0.05
0.10
0.15
0.20
Epsilon


3
2
1
Efficiency Medi
0.00
0.05
0.10
0.15
0.20
Epsilon


6
4
2
0
c
0.00
0.05
0.10
0.15
0.20
e

END

9-87

DTIC